A PHONE-DEPENDENT CONFIDENCE MEASURE FOR UTTERANCE REJECTION

Ze'ev Rivlin, Michael Cohen, Victor Abrash, and Thomas Chung
Speech Technology and Research Laboratory
SRI International
Menlo Park, California 94025
zev@speech.sri.com

ABSTRACT

An acoustic confidence measure for acceptance/rejection of recognition hypotheses for continuous speech utterances is proposed. This measure is useful for rejecting utterances that are out of domain, or contain out-of-vocabulary words or speech disfluencies. A phone-based approach is implemented so that a single global threshold can be applied to hypothesis rejection for any word sequence. Phone confidence is computed for each frame of speech as the posterior phone probability given the acoustic observation. Word sequence confidence is evaluated as the average phone confidence, either by weighting all frames equally or by normalizing by phone duration. The confidence measure is tested on a database of spoken company names. When normalized by phone duration, it achieves, in some cases with less computational expense, rejection performance comparable to a baseline system implementing a common filler-model approach. When all frames are equally weighted, performance is substantially poorer.

1. INTRODUCTION

When continuous speech recognition systems are fielded to a large community of users, especially infrequent users (e.g., callers of a telephone information service), it is common that many spoken inputs do not fall within the domain that the recognition system is designed to handle. This may be due to speech disfluencies on the part of the user (e.g., hesitations, word fragments, corrections), an incomplete language model (i.e., the system does not model all the word strings users say), or a poor understanding of the domain by the user (i.e., the user does not understand the range of inputs allowable at that point in the interaction). The ability to reject out-of-domain utterances is essential for the design of user-friendly interfaces.

A number of rejection approaches have been suggested in the past for rejection of putative hits in keyword spotting (e.g., [1, 2, 3, 4, 5, 6, 7]), for detection of out-of-vocabulary words (e.g., [8]), and for utterance rejection (e.g., [8, 9]). Some of these systems use a filler model to match non-keyword speech. A typical filler model is a set of context-independent phonetic models. Also, some systems use anti-keyword models. For example, in the digit recognition system in [9], for each digit, an anti-digit model was trained on all digits except the target digit. A central issue in all these approaches is the normalization of acoustic likelihood scores of recognition hypotheses.

We propose a phone-based confidence measure for rejecting recognition hypotheses. A recognition hypothesis (for an uttered word sequence) is rejected if its overall confidence score falls below a threshold. Two variations of the phone-based confidence measure are compared. Although we demonstrate here the application of our rejection strategy to a system *without* keyword spotting ability (i.e., the case when the only acceptable inputs are in-domain spoken word sequences unaccompanied by extraneous speech), the same strategy can be used for rejecting putative keyword hits while wordspotting.

Section 2 describes the confidence measure, Section 3 describes experimental results, and Section 4 presents conclusions and future directions.

2. PHONE-BASED CONFIDENCE MEASURE

Let $\mathbf{PH} = \{PH_1, PH_2, ..., PH_N\}$ be a Viterbi decoded sequence of phones for a spoken utterance. Let $\mathbf{O} = \{O_1, O_2, ..., O_T\}$ be the acoustic observation sequence for the utterance. Equivalently, $\mathbf{O} = \{O_{b[1]}, ..., O_{e[1]}, O_{b[2]}, ..., O_{e[2]}, ..., O_{b[N]}, ..., O_{e[N]}\}$, where $b[i]$ and $e[i]$ denote, respectively, the beginning and ending frames of the i^{th} phone. Note that $b[1] = 1$ and $e[N] = T$. Although our recognition system uses context-dependent phones, *context-independent* phones are used (for implementation reasons) to calculate the acoustic confidence measure (ACM), for which there are two variations, ACM_1 and ACM_2. ACM_1 (Equation 1) is the average per-frame log phone posterior probability. ACM_2 (Equation 2) is the average duration-normalized log phone posterior probability. The important distinction is that ACM_1 weights all *frames* equally in their contribution to the overall confidence, whereas ACM_2 weights all *phones* equally.

Equation 3 defines the posterior phone probability for O_t. In Equation 3, the local acoustic observation likelihood for a given phone, $p(O_t|PH_j)$, is computed as the maximum of the likelihood scores of the acoustic observation over all 3 states of the context-independent phone hidden Markov model (HMM). The denominator of Equation 3 is a sum over all *context-independent* phone HMMs in a phonetically tied mixture system (a system in which only HMM states that belong to allophones of the same phone share the same mixture components). Note that for a phonetically tied mixture system, the denominator is *exactly* $p(O_t)$, the unconditional likelihood of the acoustic observation, considering all *context-dependent* phone models in the system. This is *not* the case for a general genonic tied mixture system [10].

$$ACM_1(\mathbf{PH}) = \frac{1}{T}[\sum_{i=1}^{N}\sum_{t=b[i]}^{e[i]} \log P(PH_i|O_t)] \qquad (1)$$

$$ACM_2(\mathbf{PH}) = \frac{1}{N}\sum_{i=1}^{N}[\frac{1}{e[i]-b[i]+1}\sum_{t=b[i]}^{e[i]} \log P(PH_i|O_t)] \qquad (2)$$

where

$$P(PH_i|O_t) = \frac{p(O_t|PH_i)P(PH_i)}{\sum_j p(O_t|PH_j)P(PH_j)} \qquad (3)$$

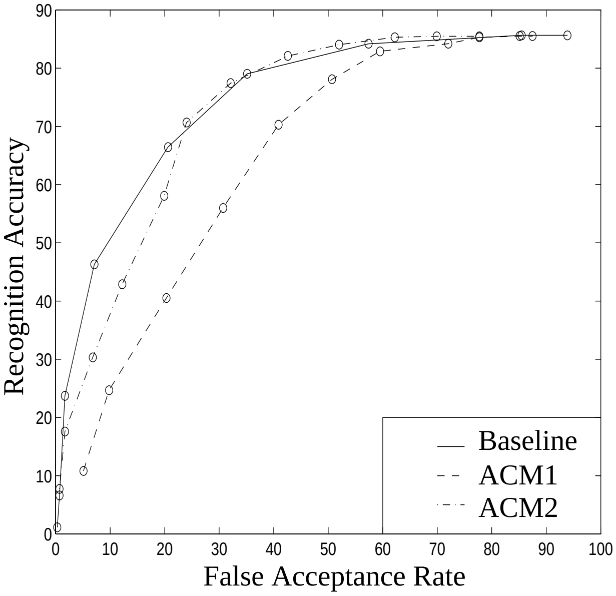
3. EXPERIMENTAL RESULTS

We conducted experiments using a database of company names spoken over the telephone to an HMM-based phonetically tied mixture continuous speech recognition system. The a priori context-independent phone probabilities, needed for Equation 3, were available from an independent database. The recognition task is to recognize which of 12,000 company names was spoken. The test set contains 916 utterances. Approximately one third of these (296 utterances) are not valid for the application, containing either no company name in the utterance, a company name not handled by the system, an acceptable company name with extraneous speech, or a subtle variant of an acceptable company name. In the test set, there are many tokens of such subtle variants as well as many in-domain company names that differ by only one or two phones. This makes the task rather challenging. The 296 utterances are what we refer to as *out-of-domain utterances.* The goal is to maximize rejection, or equivalently, to minimize false acceptance of these out-of-domain utterances. All other utterances are considered *in domain*, and should *not* be rejected.

Figure 1 shows recognition accuracy on the in-domain utterances as a function of false acceptance rate on the out-of-domain utterances. The baseline uses a rejection model implemented previously in a number of other systems (e.g., [3]), based on a filler model consisting of a set of context-independent phones. A weight is used to adjust the tradeoff between correct acceptance (i.e., not rejecting an in-domain utterance) and correct rejection performance (i.e., rejecting an out-of-domain utterance). The ACM_1 and ACM_2 systems use a threshold to determine when to reject a recognition hypothesis. For computing the hypothesis confidence via ACM_1 and ACM_2, regions recognized as non-speech were ignored.

It is clear from the figure that the confidence measure that weights all *frames* equally (ACM_1) performs significantly *worse* than that which weights all *phones* equally (ACM_2) for all false acceptance rates. One possible reason for this is the following. When the recognized word sequence shares many phones with the correct word sequence, but has several extra phones, the corresponding phone HMMs for these extra phones must be traversed across an acoustic observation region which corresponds to uttered phones that are different from those recognized. Typically, in order to get the best recognition match, these phones will have minimal duration in the Viterbi backtrace. In our system, the minimal duration is 3 frames for our 3-state phone models. Furthermore, since these recognized phones are incorrect, they typically have very poor likelihood scores. In these cases, the confidence measure more sensitive to these 3 frames of very poor likelihood scores would be able to identify the misrecognition and reject. Since ACM_2 weights phones equally, these very poor likelihood scores would have more weight. In contrast, for (ACM_1), since these phones have a minimal Viterbi duration (3 frames in our system), they would have less weight. We have started to investigate this theory and there is anecdotal evidence that it's valid.

Figure 1. Rejection Method Performance Comparison



ACM_2 achieved the same recognition accuracy as the baseline filler model for false acceptance rates greater than 22 percent. In some cases, the ACM_2 approach may be less expensive to implement than the filler-model approach. At lower false acceptance rates, the baseline model outperformed both ACM_1 and ACM_2, although the recognition accuracy for all methods was quite poor in this region.

4. CONCLUSIONS AND FUTURE DIRECTIONS

An acoustic confidence measure (ACM) for word-string hypotheses is proposed. The hypothesis confidence is evaluated as the average phone confidence. We experimented with two variations of the acoustic confidence measure, one that weights all *frames* equally (ACM_1), and one that weights all *phones* equally by normalizing for phone duration (ACM_2).

ACM_2 provided performance comparable to our baseline system, that uses a set of context-independent phones as a filler model. The ACM_2 scheme may be less expensive to implement than the filler-model approach in some cases. ACM_1 provided significantly worse performance. There is anecdotal evidence that this poorer performance is due to ACM_1's lower sensitivity to outlier, very poor likelihood scores that occur in minimal-duration phones typically indicative of a misrecognition in which the hypothesis shares many phones with, but has several more phones than the correct word sequence.

One issue that we wish to address next is normalizing the phone confidences with respect to phone model performance [11]. Also, from our comparison of ACM_1 and ACM_2, it seems it would be advantageous to incorporate durational information in confidence scoring (i.e., rather than just normalizing for duration). In addition, we wish to use *context-dependent* phone models to evaluate confidence measures in order to improve the estimation of posterior phone probabilities. Finally, we would like to apply our approach to a keyword spotting system in which we would compute word-level confidences as an average of the phone confidences for the phones making up the word. For this task, the confidence measure presented here could be used, and, with proper normalization [11], a single threshold could accommodate all keywords, eliminating the problem of determining thresholds for keywords that are uncommon.

ACKNOWLEDGMENTS

Support for this work from ARPA through the Naval Command, Control and Ocean Surveillance Center Contract N66001-94-C-6046 is gratefully acknowledged. The Government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Government funding agencies.

REFERENCES

[1] R.C. Rose and D.B. Paul, "A Hidden Markov Model Based Keyword Recognition System," *1990 IEEE ICASSP*, pp. 129-132, 1990.

[2] E. Lleida et al., "Out-of-vocabulary Word Modelling and Rejection for Keyword Spotting," *1993 Eurospeech*, pp. 1265-1268, 1993.

[3] R.C. Rose, "Discriminant Wordspotting Techniques for Rejecting Non-vocabulary Utterances in Unconstrained Speech," *1992 IEEE ICASSP*, Vol. II, pp. 105-108, 1992.

[4] M. Weintraub, "LVCSR Log-Likelihood Ratio Scoring for Keyword Spotting," *1995 IEEE ICASSP*, Vol. I, pp.297-300.

[5] H. Bourlard, B. D'hoore, and J-M Boite, "Optimizing Recognition and Rejection Performance in Wordspotting Systems," *1994 IEEE ICASSP*, Vol. I, pp. 373-376, 1994.

[6] H. Gish and K. Ng, "A Segmental Speech Model with Applications to Word Spotting," *1993 IEEE ICASSP*, Vol. II, pp. 447-450, 1993.

[7] J.R. Rohlicek et al., "Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting," *1989 IEEE ICASSP*, pp. 627-630, 1989.

[8] S.R. Young, "Detecting Misrecognitions and Out-of-vocabulary Words," *1994 IEEE ICASSP*, Vol. II, pp. 21-24.

[9] M.G. Rahim, C.H. Lee, and B-H Juang, "Robust Utterance Verification for Connected Digits Recognition," *1995 IEEE ICASSP*, pp. 285-288, 1995.

[10] V. Digalakis and H. Murveit, "Genones: Optimizing the Degree of Tying in a Large Vocabulary HMM-based Speech Recognizer," *1994 IEEE ICASSP*, Vol. I, pp. 537-540.

[11] Z. Rivlin, "A Confidence Measure for Acoustic Likelihood Scores," *Eurospeech 1995*, Vol. I, pp. 523-526, 1995.

A PHONE-DEPENDENT CONFIDENCE MEASURE FOR UTTERANCE REJECTION

Ze'ev Rivlin, Michael Cohen, Victor Abrash, and Thomas Chung
Speech Technology and Research Laboratory
SRI International
Menlo Park, California 94025
zev@speech.sri.com

An acoustic confidence measure for acceptance/rejection of recognition hypotheses for continuous speech utterances is proposed. This measure is useful for rejecting utterances that are out of domain, or contain out-of-vocabulary words or speech disfluencies. A phone-based approach is implemented so that a single global threshold can be applied to hypothesis rejection for any word sequence. Phone confidence is computed for each frame of speech as the posterior phone probability given the acoustic observation. Word sequence confidence is evaluated as the average phone confidence, either by weighting all frames equally or by normalizing by phone duration. The confidence measure is tested on a database of spoken company names. When normalized by phone duration, it achieves, in some cases with less computational expense, rejection performance comparable to a baseline system implementing a common filler-model approach. When all frames are equally weighted, performance is substantially poorer.